DETERMINATION OF SENSITIVE MEASURES OF PILOT WORKLOAD AS A FUNCTION OF THE TYPE OF PILOTING TASK

WALTER W. WIERWILLE
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

SUMMARY

The purpose of our present work, sponsored by NASA-AMES, is to examine the sensitivity, intrusion, and transferability of a variety of workload assessment techniques. The study will use four different simulated piloting tasks, emphasizing psychomotor, perceptual, mediational, and communications aspects. Pilot loading levels will be systematically adjusted. Our simulation facility is a GAT-1B that has been modified and instrumented for workload estimation techniques measurement. The flight simulator itself has three degrees of physical motion and a full complement of IFR instruments.

Recently we completed the experiment emphasizing the psychomotor aspect of flight. Instrument-rated pilots flew instrument approaches under three combined settings of the independent variable: increasing turbulance and decreasing longitudinal stability. Twenty different workload measures were taken between the outer and middle markers, only five of which showed statistically reliable changes as a function of the independent variable. Included in the five were: two rating scales, one measure of control movement activity, pulse rate, and one measure of time estimation. The results of the experiment are to some extent surprising, for they indicate that several "accepted" measures of workload are not reliably sensitive to the kinds of psychomotor load which pilots encounter.

We are currently planning the perceptual and mediational (cognitive) experiments, and we expect to have the results of these two experiments in mid-1982.

INTRODUCTION

The increasing complexity of aircraft systems and the changing roles of pilots and other aircrew personnel have resulted in the need for techniques to measure operator workload in a wide range of situations and tasks. One need only initiate a preliminary survey of the literature on operator workload assessment techniques to discover that a voluminous mass of information has accumulated rapidly in the past two decades. However, major reviews of this literature have concluded that while workload research has advanced in both scope and technology, basic questions remain to be answered for the practitioner who wishes to select workload measures for a given application (Wierwille and Williges, 1978). Hicks and Wierwille (1979) have pointed out that, in particular, the lack of information on the relative sensitivity, the degree of intrusion, and the range of transferability of individual techniques makes it difficult for a practitioner to select workload estimation techniques for a given task.

The purpose of our present work is to help fill the need for practical information. Specifically, techniques for measurement of pilot workload are being selected and compared to determine their relative sensitivity, intrusion, and transferability.

Before proceeding with further discussion and results of our experiments up to the present, it would be helpful to define the terms, sensitivity, intrusion, and transferability. Sensitivity can be defined as the relative ability of a workload estimation technique to discriminate statistically significant differences in operator loading. High sensitivity requires discriminable changes in

the score means as a function of load level and low variation of the scores about the means. Intrusion can be defined as an undesirable change in the task for which workload is being estimated, resulting from the introduction of a workload measurement technique or apparatus. And, transferability is the relative ability of a workload estimation technique to remain sensitive when being applied in situations requiring different operator behaviors or skills.

Unfortunately, there has been no definitive major effort aimed at sensitivity, intrusion, and transferability. As a result, progress in determining which workload estimation techniques should be selected for a given application has been painfully slow. When asked which techniques are sensitive in a given piloting situation, an honest workload researcher has difficulty responding. The danger is that in a given application insensitive techniques may be used. These techniques would show no substantial change in workload when in fact there is a change. Unless one knows that a technique is sensitive in a given situation, one has no assurance the evaluation of workload in an experimental situation will result in definitive conclusions.

In our work we have arbitrarily divided piloting behavior into four categories: psychomotor, perceptual, mediational, and communications. These four behaviors are those suggested by Berliner, Angell, and Shearer (1964), in their list of universal operator behaviors (See also Wierwille and Williges, 1980). Clearly, other task taxonomy categories might have been chosen. However, the Berliner, et al categories do appear to reflect the major categories of behaviors exhibited by pilots and other aircrew members.

Our evaluation of workload estimation techniques in psychomotor tasks has been completed. Results of the study will be summarized in the following sections

of this paper. More complete descriptions appear on Connor 1981 and Connor and Wierwille, 1982.

Presently, we are in the final stages of planning for the experiments emphasizing perceptual behavior and mediational behavior. In both cases pilots will fly the simulator in the simple task of maintaining heading, airspeed, and altitude. For the perceptual task they will also perform a forced-pace visual search task presented through the windscreen using an Ektagraphic display. The complexity of the search task will become the independent variable. For the mediational task, navigation problems will be presented. The problems will be forced-pace, but will not require computational aids. We expect to have the results of these two experiments in mid-1982. After these experiments have been completed we will also plan and carry out a simulated task involving communications. In all the experiments, loading level will be the independent variable, and technique scores will be the dependent variables. By conducting the experiments using this philosophy, we can obtain direct comparisons of the sensitivity of various techniques. Transferability will be evaluated by determining which (if any) techniques remain sensitive from one experiment to another . And, since primary task measures are taken for all techniques, intrusion can be determined by comparision of primary task measures with and without the other workload measurement techniques.

Clearly, the studies we are performing must necessarily be limited in scope, and they will not answer all important questions about sensitivity, intrusion, and transferability. Nevertheless we believe they will be very helpful to practitioners who must evaluate workload in realistic aircraft environments.

REVIEW OF THE EXPERIMENT ON PSYCHOMOTOR LOAD

Subjects

Six male instrument-rated pilots served as subjects in this experiment. The flight time of the subjects ranged from 500 to 2700 hours with a mean of 1300 hours.

Apparatus

The primary apparatus in this experiment was a modified flight task simulator (Singer Link, Inc., General Aviation Trainer, GAT-1B). The simulator had three degrees of freedom of motion (roll, pitch, and yaw). Translucent blinders were used to cover the windows of the simulator to reduce outside distractions and cues and to aid in the control of cockpit illumination.

Several modifications to the flight simulator were made for the experiment. These modifications permitted primary task load manipulation, secondary task operations, response measurement, and scoring. Primary task load manipulation was accomplished by changing aircraft pitch stability and random windgust disturbance level simultaneously. Three load conditions were developed: low, medium, and high, as shown in Table 1.

Table 2 provides a list of the workload measurement techniques selected for inclusion in the present study. These techniques were selected on one of two bases. First, evidence was found which indicated that the measures might be sensitive indicators of pilot workload in both simulated and operational flight. Second, previous research had shown that these measures could be useful in a variety of tasks relevant to the flight environment. A review of the twenty techniques selected can be found in Connor (1981).

Experimental Design

A complete 3 x 20 within-subject design was used for the sensitivity analysis. Load was the factor with three levels. Measurement technique (Table 2) was the factor with twenty levels.

Workload measures from different techniques were taken simultaneously on some of the data collection runs. Only those measures which were not likely to affect each other were taken simultaneously. Table 3 shows the scheme used for combining different measurement techniques for data collection. The combination of measurement techniques shown in the table was, to an extent, based on previous investigations of workload. Hicks and Wierwille's (1979) study supported the combination in condition 2. The two rating scales were administered in separate measurement conditions to prevent the ratings on one scale from biasing the ratings on the other scale. The secondary task measures were divided among several conditions because of potential intrusion and interference. Vocal measures were recorded from the two secondary tasks which required a verbal response as per Schiflett and Loikith's (1979) recommendation.

It should be noted that primary task measures were recorded on all subjects and on all data collection flights for the intrusion analysis. However, only data from measurement condition 1 were used for the sensitivity analysis of the primary task measures.

The intrusion analysis was designed to examine the effect of measurement condition, and the interaction of measurement condition with load on primary task performance. Data for all primary task measures were therefore collected for each flight performed in the six measurement conditions.

General Procedure

After receiving instructions, subjects flew nine familiarization flights in the simulator. These flights were similar, but not the same as, the data collection flights. All subjects flew the familiarization flights in the same order. Steady crosswinds were introduced for each run, and subjects were given heading corrections.

After the familiarization session, the subjects participated in three data collection sessions. The familiarization session and each data collection session were held on a different day.

Each data collection session consisted of two sets of a warm-up practice flight and three data collection flights. The practice flight was the same as the first data collection flight. Since the data collection flights were counterbalanced, equal amounts of practice were provided for the low, medium, and high load conditions. The data collection flights also contained steady crosswind conditions, for which the subject was given heading corrections. The purpose of introducing steady crosswinds was to disguise the load conditions, thereby requiring subjects to fly each flight as a separate entity.

Flight Task Procedures

The flight task in this experiment was an Instrument Landing Systems (ILS) approach to the Seaport Beach runway (29L) which is instrumented in the Singer Link GAT-1B aircraft simulator. Prior to the beginning of a flight, the simulated aircraft was positioned 5 miles outbound from the Seaport Beach outer marker on the 108 degree radial, heading into the wind. When ready to begin, the experimenter informed the subject of the wind direction and speed, and gave him a heading cor-

rection for the crosswind. When contacted by the experimenter, the subject took off and climbed to 2000 feet. The subject then flew directly to the outer marker by following the localizer at 100 miles per hour until the glide slope was intercepted. Upon interception of the glide slope, the subject reduced airspeed to 80 miles per hour and proceeded down the glide slope while following the localizer to a landing. Data were recorded between the outer and middle markers. For the opinion measures, subjects gave ratings for the flight segment between the outer and middle markers immediately after landing and parking the simulated aircraft.

Results

Sensitivity Analysis

The computed scores for each technique were first converted to Z-scores (normalized scores) so that technique measure units would not affect the sensitivity analysis. Subsequently, an overall analysis of variance was performed on the scores. Since Z-scores were used, a technique main effect was not possible. A significant main effect of load was found, $F(2,10) = 5.34$, $p < 0.0001$, and a significant load by technique interaction was found, $F(38,190) = 2.76$, $p \leq 0.05$.

The load by technique interaction indicated that the measurement techniques were differentially sensitive to load. Therefore, individual ANOVAs were used to isolate the sensitive techniques.

The individual ANOVAs indicated that five of the twenty measures were sensitive. They were the Cooper-Harper scale $F(2,10) = 16.39$, $p = 0.0007$; the Workload Compensation-Interference/Technical Effectiveness (WCI/TE) scale, $F(2, 10) = 31.15$, $p < 0.0001$; the time estimation standard deviation, $F(2,10) = 5.69$,

$p = 0.022$; the pulse rate mean, $F(2,10) = 8.89$, $p = 0.006$; and the control movements measure, $F(2,10) = 33.84$, $p < 0.0001$. The normalized means for each technique are plotted in Figures 1 through 5 as a function of load.

Newman-Keuls comparisons were then performed on the normalized means of the sensitive measures. The comparisons included low vs. medium, medium vs. high, and low vs. high load conditions. Results indicated that all differences were significant at $p < 0.05$, except for pulse-rate mean (low vs. medium and medium vs. high) and time estimation standard deviation (low vs. high).

A logical classification of techniques based on demonstrated sensitivity was generated from an examination of the Newman-Keuls comparisons, as shown in Table 4. Techniques which demonstrated sensitivity to all pairs of load conditions (i.e., low vs. medium, medium vs. high, and low vs. high) were included in class I. These measures are preferred over other techniques which demonstrated only partial sensitivity, or no sensitivity in the present study. Techniques which showed sensitivity to some differences in load conditions (but not all) were included in class II. These measures are less preferred than class I techniques, but are more preferred than class III techniques. Class III techniques did not demonstrate sensitivity to load in the present study. This class includes all techniques except those in class I and class II.

One possible reason that only five of the twenty workload assessment techniques demonstrated sensitivity in the present study is that the other techniques simply required a greater number of subjects to show a significant effect of load. It is possible to estimate the sample size required to detect a reliable load effect for a given workload assessment technique at specified levels of significance and power. These calculations were performed for those techniques which did not demonstrate

sensitivity in the present study, to provide an indication of the practical costs of achieving statistical significance. The procedure used for estimating the sample size required for finding sensitivity is described by Bowker and Lieberman (1959). Sample sizes were estimated for a significance level of 0.05 and for a power of approximately 0.80. The results of these estimates are presented in Table 5.

Intrusion Analysis

The equipment and procedures used for some workload assessment techniques may interfere with performance on the primary (flight) task. In the present experiment, data for the twenty measurement techniques were recorded in six measurement conditions as shown in Table 3. These six measurement conditions differed in the equipment and procedures used for data collection. The purpose of the intrusion analysis was to examine the effect of these measurement conditions on primary task performance.

The equipment and procedures used in measurement condition 1 were assumed to be unobtrusive to primary task performance. Primary task performance in this condition was therefore used as a standard of comparison for primary task performance on the other five measurement conditions. The measures of primary task performance which were used for these comparisons included scores on localizer rms error, glide slope rms error, and control movements per second.

A multivariate analysis of variance (MANOVA) was performed to examine the effect of condition, load, and the interaction of condition and load on the primary task measures. Only the main effect of load was found to be significant F (2,10) = 9.42, p = 0.0002. Because there was no significant effect of condition nor sig-

nificant interaction of condition with load, it can be concluded that the physiological measuring equipment and the secondary tasks did not significantly affect pilot performance in terms of the three primary task variables.

Conclusions

This study has shown that five measures of workload estimation were sensitive indicators of load in a piloting task that is predominantly psychomotor in nature. Another fifteen measures, believed to be "good" measures of workload, showed no reliable effect. The main conclusion that must be drawn from the study is that few measures are sensitive to psychomotor load.

Of the five techniques demonstrating sensitivity, only three exhibited monotonic score increases with load as well as statistically reliable differences between all pairs of load levels. Consequently, only the three meet all criteria for sensitivity to psychomotor load. These class I techniques are the ones that are recommended for measurement of psychomotor load:

Cooper/Harper ratings,

WCI/TE ratings, and

Control movements per second.

The other two techniques showed sensitivity to psychomotor load, but did not discriminate between all pairs of load levels. These class II techniques are:

Time estimation standard deviation, and

Pulse rate mean.

These measures would be helpful in evaluating psychomotor load, but they should not be relied on exclusively. At least one class I technique should also be used in conjunction with these measures.

It is worth noting that only two opinion measures were taken in the present experiment, and both proved sensitive. This suggests that well-designed rating scales are among the best of techniques for evaluating psychomotor load. In regard to the primary task measures, the control movements measure alone was sensitive. However, this measure is also the only primary task measure which reflected "strategy" of the pilot. Consequently, one could speculate that selecting a primary task measure that reflects strategy will most likely result in good sensitivity.

Fifteen (techniques) measures showed no reliable change as a function of load. When these fifteen measures were subjected to a power analysis to determine sample size, the number of subjects required ranged from 12 to well over 100 (Table 5). One can only conclude that at best the fifteen measures, as taken, are much less sensitive to psychomotor load than the five appearing in Classes I and II. Of course, there is always the possibility that the measures would be sensitive to loading along other dimensions of human performance, such as psychomotor tasks of a different nature, or mediational or cognitive tasks, for example.

In regard to intrusion, this experiment showed that no significant interference occurred for the physiological measures or for the secondary task measures. Performance as measured using three primary (flight) task measures showed no reliable changes as a function of addition of these measures.

In general, the results of the experiment show that there are wide variations in the sensitivity of workload estimation measures. Great care must be taken in selecting measures for a given experiment. Otherwise, it is possible that no changes in workload will be found, when indeed there are changes.

References

Berliner, C., Angell, D., and Shearer, D. J. Behaviors, measures, and instruments for performance evaluation in simulated environments. Paper presented at the Symposium and Workshop on the Quantification of Human Performance, Albuquerque, New Mexico, August, 1964.

Bowker, A. H. and Lieberman, G. J. Engineering statistics. New Jersey: Prentice-Hall, Inc., 1959.

Connor, S. A. A comparison of pilot workload assessment techniques using a psychomotor task in a moving base aircraft simulator. Master's Thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, October, 1981.

Connor, S. A. and Wierwille, W. W. Comparative evaluation of twenty pilot workload assessment measures using a psychomotor task in a moving base simulator. Moffett Field, CA: NASA-Ames Research Center, (Forthcoming report).

Hicks, T. G. and Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving base driving simulator. Human Factors, 1979, 21, 129-143.

Schiflett, S. G. and Loikith, G. J. Voice stress as a measure of operator workload. Patuxent River, Maryland: Naval Air Test Center, Technical Memorandom TM 79-3 SY, December 31, 1979.

Wierwille, W. W. and Williges, R. C. Survey and analysis of operator workload assessment techniques. Blacksburg, Virginia: Systemetrics, Inc. Report No. S-78-101, September, 1978.

Wierwille, W. W. and Williges, B. H. An annotated bibliography on operator mental workload assessment. Patuxent River, Maryland: Naval Air Test Center Report No. SY-27R-80, March, 1980.

Acknowledgements

The author wishes to thank Mr. Sidney A. Connor, who carried out the psychomotor workload experiment, and Mrs. Sandra Hart, NASA-Ames Research Center, for helpful technical suggestions.

TABLE 1

Primary Task Load Conditions

	LOAD CONDITION		
	Low	Medium	High
RANDOM GUST LEVEL	Low	Medium	High
Estimated Std. Dev. (mph)	0	2.7	5.9
PITCH STABILITY	High	Medium	Low
a. Control input to pitch rate output equivalent gain (degrees/s per % of control range)	0.522	3.560	7.83
b. Control input to pitch rate output equivalent time constant (s)	0.097	0.660	1.45

TABLE 2

Workload Assessment Techniques Which Were Tested in the Present Experiment

---

OPINION
1. Cooper-Harper Scale
2. WCI/TE Scale

SPARE MENTAL CAPACITY
3. Digit Shadowing (% errors)
4. Memory Scanning (Mean time)
5. Mental Arithmetic (% errors)
6. Time Estimation Mean (Seconds)
7. Time Estimation Standard Deviation (Seconds)
8. Time Estimation Absolute Error (Seconds)
9. Time Estimation RMS error (Seconds)

PHYSIOLOGICAL
10. Pulse Rate Mean (Pulses per minute)
11. Pulse Rate Variability (Pulses per minute)
12. Respiration Rate (Breath cycles per minute)
13. Pupil Diameter (Normalized units)
14. Voice Pattern (Digit Shadowing Task)
15. Voice Pattern (Mental Arithmetic Task)

EYE BEHAVIOR
16. Eye Transition Frequency (Transitions per minute)
17. Eye Blink Frequency (Blinks per minute)

PRIMARY TASK
18. Localizer RMS Angular Position Error (Degrees)
19. Glide Slope RMS Angular Position Error (Degrees)
20. Control Movements per second
    (Aileron + Elevator + Rudder)

TABLE 3

Combination of Measurement Techniques
for Data Collection

Measurement Condition	Measurement Techniques
1.	Cooper-Harper Scale Pupil Diameter Eye Transition Frequency Eye Blink Frequency Localizer RMS Error Glide Slope RMS Error Control Movements
2.	WCI/TE Scale Pulse Rate Mean Pulse Rate Variability Respiration Rate
3.	Digit Shadowing Voice Pattern
4.	Memory Scanning
5.	Mental Arithmetic Voice Pattern
6.	Time Estimation (Mean) (Std. Dev.) (Abs. Error) (RMS Error)

TABLE 4

Logical Classification of Techniques
Based on Demonstrated Sensitivity

Class I: Complete Sensitivity Demonstrated
- Cooper-Harper Scale
- WCI/TE Scale
- Control Movements/Unit Time

Class II: Some Sensitivity Demonstrated
- Time Estimation Standard Deviation*
- Pulse Rate Mean **

Class III: Sensitivity Not Demonstrated
- All Other Techniques (See Table 5)

*Double valued function
**Limited sensitivity

TABLE 5

Estimated Sample Sizes Required for Achieving a Significant
Load Effect for Techniques not Demonstrating Sensitivity

Technique	Estimated Sample Size
SPARE MENTAL CAPACITY	
Digit Shadowing	18
Memory Scanning	>100
Mental Arithmetic	25
Time Estimation (Mean)	53
Time Estimation (Abs. Error)	>100
Time Estimation (RMS Error)	85
PHYSIOLOGICAL	
Pulse Rate Variability	45
Respiration Rate	15
Pupil Diameter	>100
Speech Pattern (D. Shadow.)	28
Speech Pattern (M. Arith.)	>100
EYE BEHAVIOR	
Eye Transition Frequency	42
Eye Blink Frequency	25
PRIMARY TASK	
Localizer RMS Error	12
Glide Slope RMS Error	41

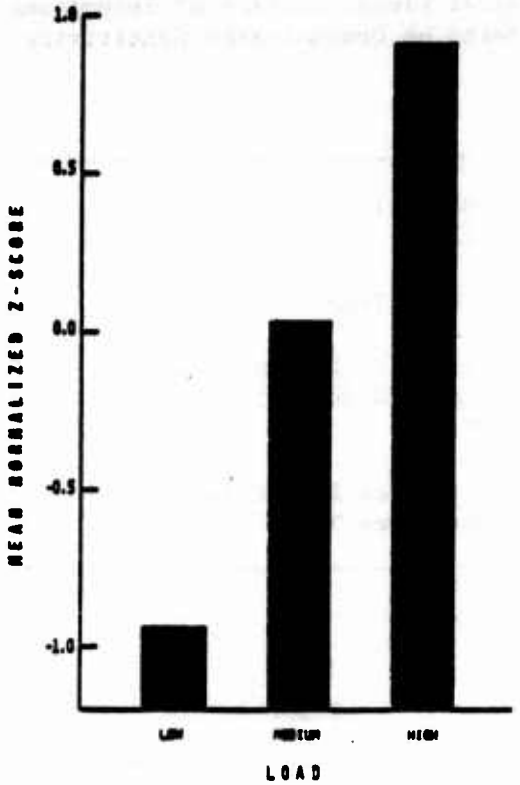

Figure 1. Mean normalized scores for the Cooper-Harper rating scale measure plotted as a function of load.

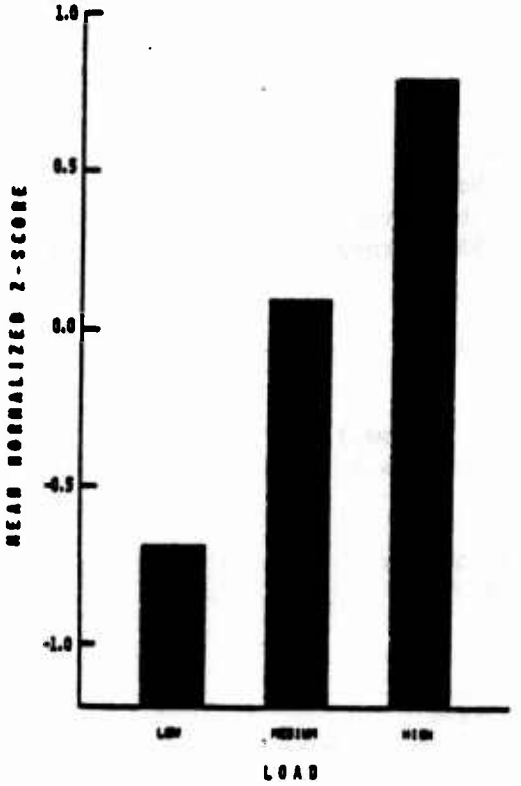

Figure 2. Mean normalized scores for the WCI/TE rating scale measure plotted as a function of load.

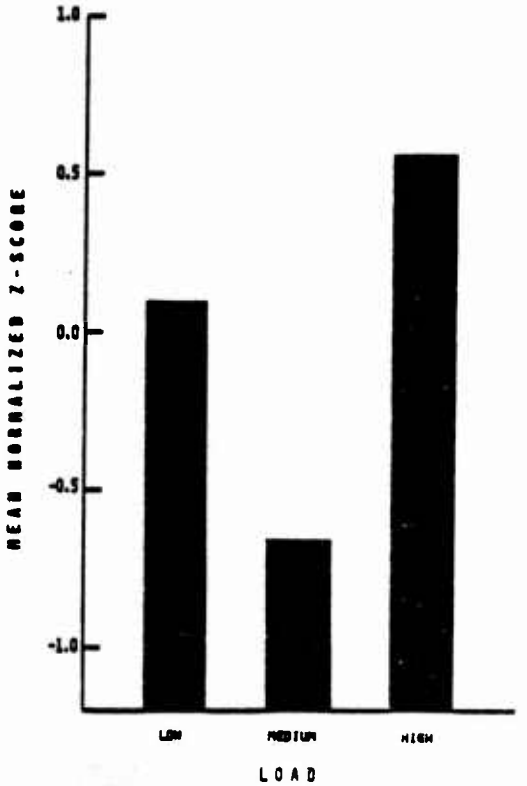

Figure 3. Mean normalized scores for the time estimation standard deviation measure plotted as a function of load.

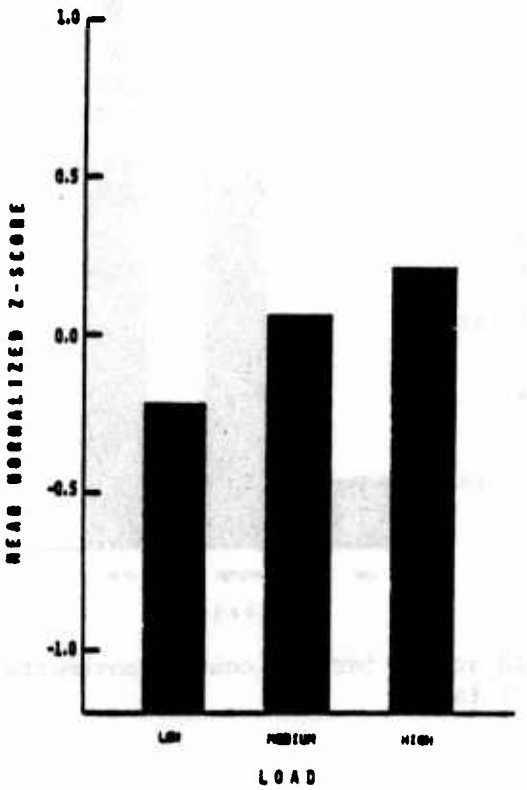

Figure 4. Mean normalized scores for the pulse rate mean measure plotted as a function of load.

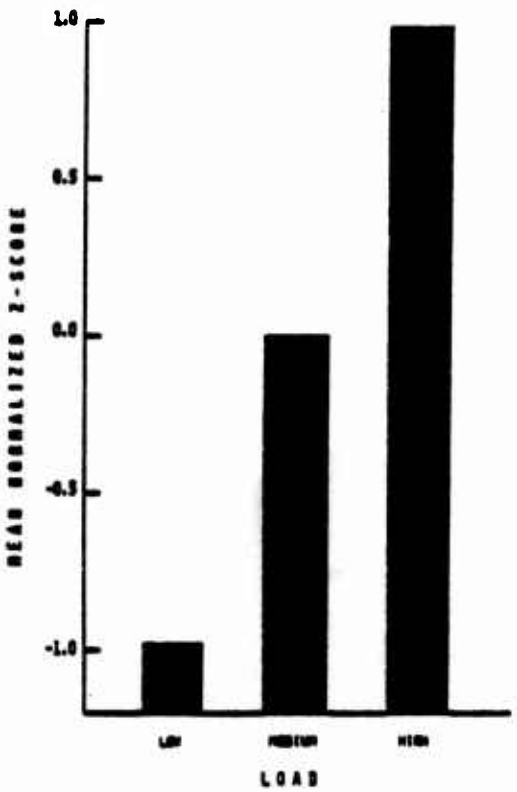

Figure 5. Mean normalized scores for the control movements measure plotted as a function of load.